Title
Genome sequence of the model rice variety KitaakeX.

Journal
BMC genomics, 20(1)

ISSN
1471-2164

Authors
Jain, Rashmi
Jenkins, Jerry
Shu, Shengqiang
et al.

Publication Date
2019-11-01

DOI
10.1186/s12864-019-6262-4

Peer reviewed

Jain et al. BMC Genomics (2019) 20:905
https://doi.org/10.1186/s12864-019-6262-4

BMC Genomics

RESEARCH ARTICLE Open Access

Genome sequence of the model rice variety KitaakeX

Rashmi Jain[1,2], Jerry Jenkins[3,4], Shengqiang Shu[3], Mawsheng Chern[1,2], Joel A. Martin[3], Dario Copetti[7,10,11], Phat Q. Duong[1,2], Nikki T. Pham[1], David A. Kudrna[7,8], Jayson Talag[7,8], Wendy S. Schackwitz[3], Anna M. Lipzen[3], David Dilworth[3], Diane Bauer[3], Jane Grimwood[3,4], Catherine R. Nelson[1], Feng Xing[6], Weibo Xie[6], Kerrie W. Barry[3], Rod A. Wing[7,8,9], Jeremy Schmutz[3,4], Guotian Li[1,2,5*] and Pamela C. Ronald[1,2*]

Abstract

Background: The availability of thousands of complete rice genome sequences from diverse varieties and accessions has laid the foundation for in-depth exploration of the rice genome. One drawback to these collections is that most of these rice varieties have long life cycles, and/or low transformation efficiencies, which limits their usefulness as model organisms for functional genomics studies. In contrast, the rice variety Kitaake has a rapid life cycle (9 weeks seed to seed) and is easy to transform and propagate. For these reasons, Kitaake has emerged as a model for studies of diverse monocotyledonous species.

Results: Here, we report the de novo genome sequencing and analysis of *Oryza sativa ssp. japonica* variety KitaakeX, a Kitaake plant carrying the rice XA21 immune receptor. Our KitaakeX sequence assembly contains 377.6 Mb, consisting of 33 scaffolds (476 contigs) with a contig N50 of 1.4 Mb. Complementing the assembly are detailed gene annotations of 35,594 protein coding genes. We identified 331,335 genomic variations between KitaakeX and Nipponbare (ssp. *japonica*), and 2,785,991 variations between KitaakeX and Zhenshan97 (ssp. *indica*). We also compared Kitaake resequencing reads to the KitaakeX assembly and identified 219 small variations. The high-quality genome of the model rice plant KitaakeX will accelerate rice functional genomics.

Conclusions: The high quality, de novo assembly of the KitaakeX genome will serve as a useful reference genome for rice and will accelerate functional genomics studies of rice and other species.

Keywords: Rice, Kitaake, KitaakeX, XA21 immune receptor, Whole genome sequence, De novo genome assembly, Nipponbare, Zhenshan97

Background

Rice (*Oryza sativa*) provides food for more than half of the world's population [1] and also serves as a model for studies of other monocotyledonous species. Cultivated rice contains two major types of *O. sativa*, the *O. sativa indica/Xian* group and the *O. sativa japonica/Geng* group. Using genomic markers, two additional minor types have been recognized, the circum-Aus group and the circum-Basmati group [2]. More than 3000 rice varieties and species have been sequenced, including Nipponbare [3], 93–11 [4], DJ 123, IR64 [5], Zhenshan97, Minghui 63 [6], Shuhui498 [7], *Oryza glaberrima* [8, 2]. The availability of these genomes has laid a strong foundation for basic rice research and breeding [2]. However, the use of these sequenced varieties for functional genomics analyses is limited by their long life cycles or low transformation efficiencies. For example, it takes up to 6 months for Nipponbare to produce seeds under winter conditions. The Indica varieties typically have relatively low transformation efficiencies [9].

The Kitaake cultivar (ssp. *japonica*), which originated at the northern limit of rice cultivation in Hokkaido, Japan [10], has emerged as a model for rice research [9]. Kitaake is insensitive to day length, easy to propagate, relatively cold tolerant, short in stature and completes its life cycle in about 9 weeks [9, 11]. These properties

* Correspondence: li4@mail.hzau.edu.cn; pcronald@ucdavis.edu
[1]Department of Plant Pathology and the Genome Center, University of California, One Shields Avenue, Davis, CA 95616, USA
Full list of author information is available at the end of the article

make it easy to cultivate under typical greenhouse conditions. Kitaake is also highly amenable to transformation [12]. Several hundred genes have been overexpressed or silenced in KitaakeX [12]. The transformation efficiency of Kitaake is comparable to that of that Dongjin, a cultivar that historically transforms well [9]. Kitaake has been used to establish multiple mutant populations, including an RNAi mutant collection [13], T-DNA insertion collections [9, 14], and a whole-genome sequenced mutant population of KitaakeX, a Kitaake variety carrying the *Xa21* immune receptor gene (formerly called X.Kitaake) [15, 16]. Kitaake has been used to explore diverse aspects of rice biology, including flowering time [17], disease resistance [18–20], small RNA biology [21], and the CRISPR-Cas9 and TALEN technologies [22, 23].

The unavailability of the Kitaake genome sequence has posed an obstacle to the use of Kitaake in rice research. For example, analysis of a fast-neutron (FN) induced mutant population in KitaakeX, a Kitaake plant carrying the rice XA21 gene [15], required the use of Nipponbare (ssp. *japonica)* as the reference genome. Additionally, CRISPR/Cas9 guide RNAs cannot be accurately designed for Kitaake without a complete sequence. To address these issues, we assembled a high-quality genome sequence of KitaakeX, compared its genome to the genomes of rice varieties Nipponbare and Zhenshan97 (ssp. *indica*), and identified genomic variations. The XA21 gene confers resistance to the bacterial pathogen, *Xanthomonas oryzae pv. oryzae,* making KItaakeX a model for studies of infectious disease [16].

Results

KitaakeX flowers significantly earlier than other sequenced rice varieties

Kitaake has long been recognized as a rapid life-cycle variety [12], but it has yet to be systematically compared to other rice varieties. We compared the flowering time of KitaakeX with other sequenced rice varieties under long-day conditions (14 h light/10 h dark). Consistent with other studies, we found that KitaakeX flowers much earlier than other varieties (Fig. 1a, b), heading at 54 days after germination. Other rice varieties Nipponbare, 93–11 (ssp. *indica*), IR64 (ssp. *indica*), Zhenshan 97, Minghui 63 (ssp. *indica*), and Kasalath (aus rice cultivar) start heading at 134, 99, 107, 79, 125, and 84 days after germination, respectively (Fig. 1b).

We next assessed how KitaakeX is related to other rice varieties using a phylogenetic approach based on the rice population structure and diversity published for 3010 varieties [2]. The 3010 sequenced accessions were classified into nine subpopulations, most of which could be connected to geographical origins. The phylogenetic tree reveals that KitaakeX and Nipponbare are closely related within the same subpopulation (Fig. 1c).

Genome sequencing and assembly

To obtain a high-quality, de novo genome assembly, we sequenced the KitaakeX genome using a strategy that combines short-read and long-read sequencing. Sequencing reads were collected using Illumina, 10x Genomics, PACBIO, and Sanger platforms at the Joint Genome Institute (JGI) and the HudsonAlpha Institute. The current release is version 3.0, which is a combination of a MECAT (Mapping, Error Correction and de novo Assembly Tools) PACBIO based assembly and an Illumina sequenced 10x genomics SuperNova assembly. The assembled sequence contains 377.6 Mb, consisting of 33 scaffolds (476 contigs) with a contig N50 of 1.4 Mb, covering a total of 99.67% of assembled bases in chromosomes (Table 1.a).

We assessed the quality of the KitaakeX assembly for sequence completeness and accuracy. Completeness of the assembly was assessed by aligning the 34,651 annotated genes from the v7.0 Nipponbare to the KitaakeX assembly using BLAT [24]. The alignments indicate that 98.94% (34,285 of genes) genes completely aligned to the KitaakeX assembly, 0.75% (259 genes) partially aligned, and 0.31% (107 genes) were not detected. A bacterial artificial chromosome (BAC) library was constructed and a set of 346 BAC clones (9.2x clone coverage) was sequenced using PACBIO

Table 1 Summary of the KitaakeX genome assembly and annotation

a. Genome characteristics and assembly		
	Estimated genome size	409.5 Mb
	Assembled contigs size	377.6 Mb
	Contig N50	1.4 Mb
	Longest contig	8.6 Mb
	Assembled scaffolds	381.6 Mb
	Scaffold N50	30.3 Mb
	Longest scaffold	44.3 Mb
	GC content	43.7%
b. Transposable elements		
	Retrotransposons size	89.6 Mb
	DNA transposons size	32.6 Mb
	Total size of transposable elements	122.2 Mb
c. Genome annotation		
	Number of protein-coding genes	35,594
	Complete BUSCOs	99.0%
	Average transcript length	1874 bp
	Average coding sequence length	1222 bp
	Number of functionally annotated genes	33,226

See method section for calculations; N50 = minimum sequence length needed to cover 50% of the genome; *BUSCOs* Benchmarking Universal Single-Copy Orthologs score

sequencing. A range of variants was detected by comparing the BAC clones to the assembly. Alignments were of high quality (< 0.1% of error) in 271 clones (Additional file 1: Figure S13). Sixty BACs indicate a higher error rate (0.45% of error) due mainly to their placement in repetitive regions (Additional file 1: Figure S14). Fifteen BAC clones indicate a rearrangement (10 clones) or a putative overlap on adjacent contigs (5 clones) (Additional file 1: Figure S15). The overall error rate in the BAC clones is 0.09%, indicating the high quality of this assembly (for detailed information, see Additional file 1).

Genome annotation

We predicted 35,594 protein-coding genes in the KitaakeX genome (Table 1.c, Additional file 2: Table S12), representing 31.5% genic space of the assembled genome size (Table 1). There is some transcriptome support for 89.5% (31,854/35,594) of the KitaakeX genes, and 81.6% (29,039/35,594) genes are fully supported by the transcriptome (Additional file 2: Table S11). The predicted protein-coding genes are distributed unevenly across each chromosome; gene density tends to be higher toward chromosome ends (Fig. 2f). The average GC content of the genome is 43.7% (Fig. 2e, Table 1.a).

To assess the quality of the annotation of KitaakeX genes, we compared the KitaakeX annotation to those of other completed rice genomes using the BUSCO v2 method, which is based on a set of 1440 conserved plant genes. The results confirm 99.0% completeness of the KitaakeX genome annotation (Table 1, Additional file 2: Table S7). To further evaluate the quality of the annotation, we studied the extent of conservation of functional genes in KitaakeX. We selected 291 genes (Additional file 3: Table S13) from three pathways associated with stress resistance, flowering time and response to

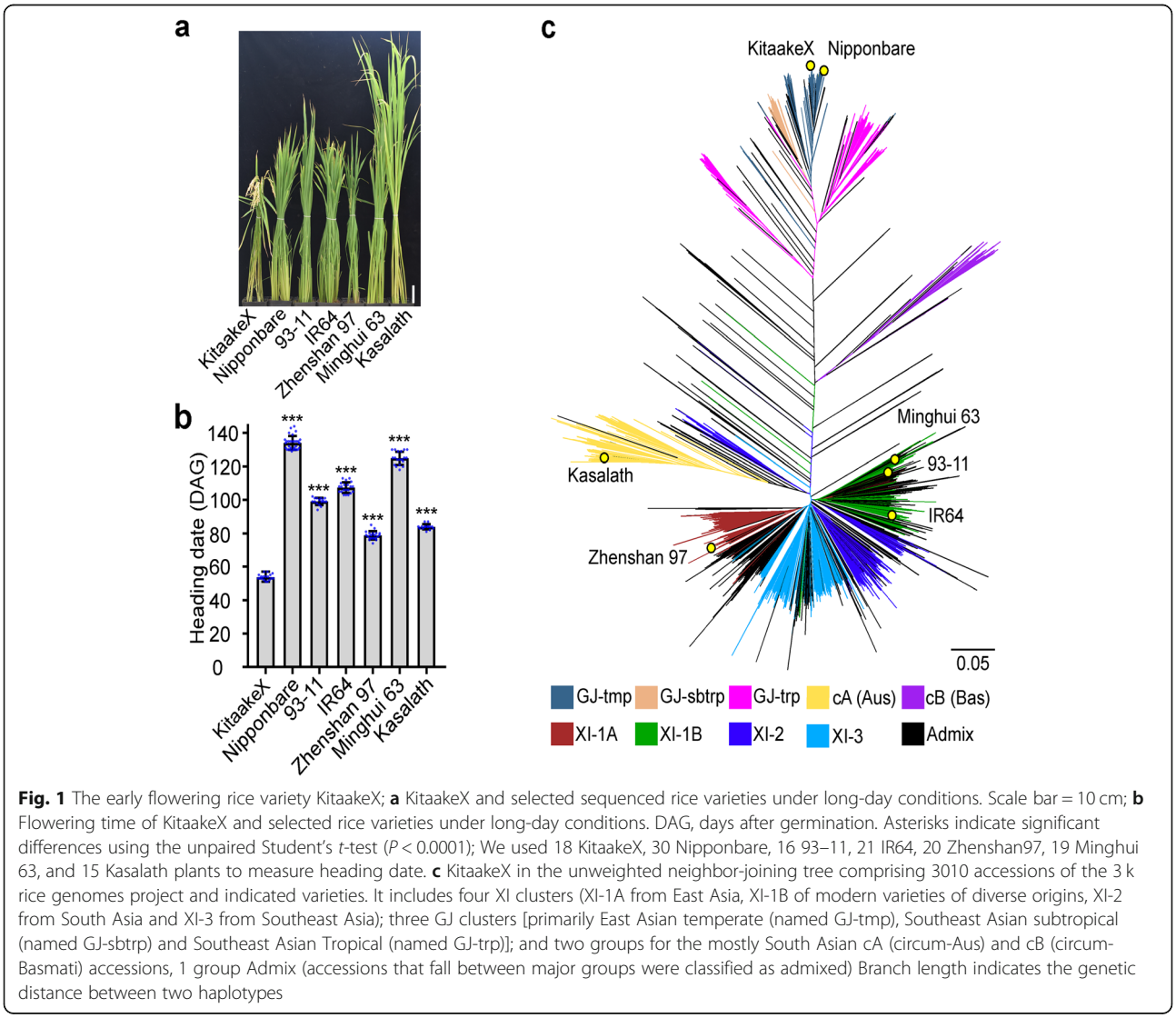

Fig. 1 The early flowering rice variety KitaakeX; **a** KitaakeX and selected sequenced rice varieties under long-day conditions. Scale bar = 10 cm; **b** Flowering time of KitaakeX and selected rice varieties under long-day conditions. DAG, days after germination. Asterisks indicate significant differences using the unpaired Student's *t*-test ($P < 0.0001$); We used 18 KitaakeX, 30 Nipponbare, 16 93–11, 21 IR64, 20 Zhenshan97, 19 Minghui 63, and 15 Kasalath plants to measure heading date. **c** KitaakeX in the unweighted neighbor-joining tree comprising 3010 accessions of the 3 k rice genomes project and indicated varieties. It includes four XI clusters (XI-1A from East Asia, XI-1B of modern varieties of diverse origins, XI-2 from South Asia and XI-3 from Southeast Asia); three GJ clusters [primarily East Asian temperate (named GJ-tmp), Southeast Asian subtropical (named GJ-sbtrp) and Southeast Asian Tropical (named GJ-trp)]; and two groups for the mostly South Asian cA (circum-Aus) and cB (circum-Basmati) accessions, 1 group Admix (accessions that fall between major groups were classified as admixed) Branch length indicates the genetic distance between two haplotypes

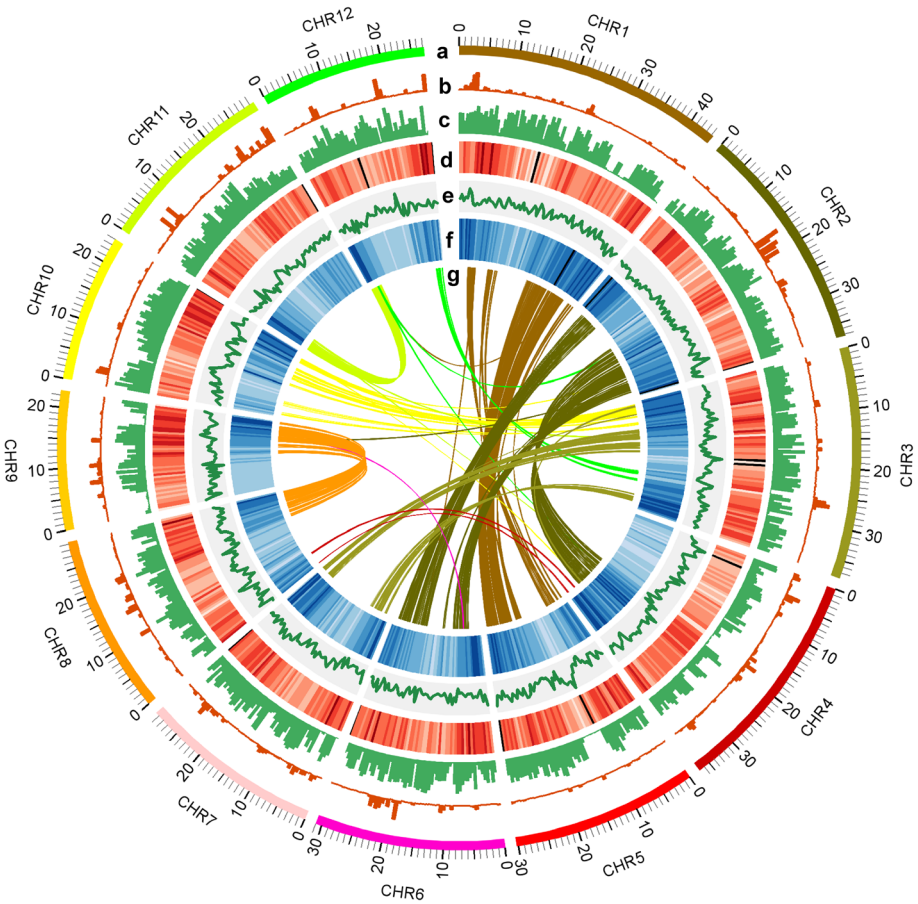

Fig. 2 Genome wide analysis of KitaakeX genome and its comparison with other rice varieties; **a** Circles indicate the 12 KitaakeX chromosomes represented on a Mb scale; **b**, **c** SNPs and InDels between KitaakeX and Nipponbare (**b**) and KitaakeX and Zhenshan97 (**c**); **d** Repeat density; **e** GC content; **f** Gene density; **g** Homologous genes in the KitaakeX genome. Window size used in the circles is 500 kb

light [8], and then searched for orthologous genes in the KitaakeX genome. We found that 275 of 291 (94.5%) of the selected KitaakeX genes show greater than 90% identity with the corresponding Nipponbare genes at the protein level. Twenty-three out of the 291 show 100% identity at the nucleotide level but not at the protein level. Of these 23 genes, the KitaakeX gene model for 16 genes has better transcriptomic evidence than the Nipponbare gene model. One of the 291 KitaakeX genes is slightly shorter than its Nipponbare ortholog due to an alternative transcript (Additional file 3: Table S13). These results indicate the high quality of the annotation, and conservation between the KitaakeX and Nipponbare *japonica* rice varieties.

Using SynMap, we identified 2469 pairs of colinear genes (88 blocks) in the KitaakeX genome (Fig. 2g). These results correlate with already published findings [25]. We used RepeatMaker and Blaster to identify transposable elements (TEs) in the KitaakeX genome, and identified 122.2 Mb of sequence corresponding to TEs (32.0% of the genome). DNA transposons account for ~ 33 Mb; retrotransposons account for ~ 90 Mb. The TEs belong mostly to the Gypsy and Copia retroelement families, and account for 23% of the genome (Additional file 2: Table S8), as is true in the Nipponbare and Zhenshan97 genomes [6].

Genomic variations between KitaakeX and other rice varieties

We compared the genome of KitaakeX to the Nipponbare and Zhenshan97 genomes to detect genomic variations, including single nucleotide polymorphisms (SNPs), insertions and deletions under 30 bp (InDels), presence/absence variations (PAVs), and inversions using MUMmer [26]. We found 331,335 variations between KitaakeX and Nipponbare (Additional file 4), and nearly 10 times as many (2,785,991) variations between KitaakeX and Zhenshan97 (Additional file 5). There are 253,295 SNPs and 75,183 InDels between KitaakeX and Nipponbare, and 2,328,319 SNPs and 442,962 InDels between KitaakeX and Zhenshan97 (Additional files 6 and Additional file 2: Table S3).

With respect to SNPs in both intersubspecies (*japonica* vs. *indica*) as well as intrasubspecies (*japonica* vs. *japonica*) comparisons, transitions (Tss) (G -> A and C -> T) are about twice as abundant as transversions (Tvs) (G -> C and C -> G) (Additional file 2: Table S10). Genomic variations between KitaakeX and Nipponbare are highly concentrated in some genomic regions (Fig. 2b), but variations between KitaakeX and Zhenshan97 are spread evenly through the genome (Fig. 2c). Intersubspecies genomic variations, then, are much more extensive than intrasubspecies variations. We also detected multiple genomic inversions using comparative genomics (Additional files 4 and 5).

For variations occurring in the genic regions, we found that single-base and 3 bp (without frame shift) InDels are much more abundant than others (Additional file 7: Figure S16a), suggesting that these genetic variations have been functionally selected. We carried out detailed analysis of gene structure alterations that exist as a consequence of SNPs and InDels between KitaakeX and Nipponbare and Kitaake and Zhenshan97. Between KitaakeX and Nipponbare, we identified 2092 frameshifts, 78 changes affecting splice-site acceptors, 71 changes affecting splice-site donors, 19 lost start codons, 161 gained stop codons, and 15 lost stop codons. In the comparison of KitaakeX to Zhenshan97, 6809 unique genes in KitaakeX are affected by 8640 frameshifts (Additional file 7: Figure S16b), 531 changes affecting splice-site acceptors, 530 changes affecting splice-site donors, 185 lost start codons, 902 gained stop codons and 269 lost stop codons (Additional file 7: Figure S16b).

Based on PAV analysis, we identified 456 loci that are specific to KitaakeX (Additional file 4) compared with Nipponbare. Pfam analysis of KitaakeX-specific regions revealed 275 proteins. Out of these 275 genes, 148 genes are from 19 different gene families with more than 2 genes in those regions. These gene families include protein kinases, leucine-rich repeat proteins, NB-ARC domain-containing proteins, F-box domain containing proteins, protein tyrosine kinases, Myb/SANt-like DNA binding domain proteins, transferase family proteins, xylanase inhibitor C-terminal protein, and plant proteins of unknown function (Additional file 7: Figure S16c). We identified 4589 loci specific to KitaakeX compared with Zhenshan97 (Additional file 5).

We also compared our de novo assembly of KitaakeX genome with Kitaake resequencing reads using an established pipeline [15]. This analysis revealed 219 small variations (200 SNPs and 19 INDELs) between the two genomes (Additional file 8). These variations affect 9 genes in KitaakeX besides the Ubi-*Xa21* transgene, including the selectable marker encoding a hygromycin B phosphotransferase on chromosome 6 (Additional file 8, Additional file 9: Figure S17).

Discussion

In 2005 the Nipponbare genome was sequenced and annotated to a high-quality level (International Rice Genome Sequencing and Sasaki 2005). Since that time, it has served as a reference genome for many rice genomic studies [27]. Despite its use, the long life cycle of Nipponbare makes it time-consuming for most genetic analyses.

Here we report the de novo assembly and annotation of KitaakeX, an early-flowering rice variety with a rapid life cycle that is easy to propagate under greenhouse conditions. We predict that KitaakeX contains 35,594 protein-coding genes, comparable to the published genomes (39,045 for Nipponbare and 34,610 for Zhenshan97) (Additional file 4 and Additional file 5). The availability of a high-quality genome and annotation for KitaakeX will be useful for associating traits of interest with genetic variations, and for identifying the genes controlling those traits.

We identified 219 SNPs and InDels between the KitaakeX and Kitaake genomes. These variations may have resulted from somatic mutations that arose during tissue culture and regeneration, or they may be spontaneous mutations [28]. For rice, 150 mutations are typically induced during tissue culture and 41 mutations occur spontaneously per three generations [28]. These numbers are consistent with the independent propagation of KitaakeX and Kitaake over approximately 10 generations in the greenhouse.

The KitaakeX genome will be useful for variety of studies. For example, we recently published the whole genome sequences of 1504 FN-mutated KitaakeX rice lines. Mutations were identified by aligning reads of the KitaakeX mutants to the Nipponbare reference genome [15]. On average, 97% of the Nipponbare genome is covered by the KitaakeX reads. However, in some regions, the KitaakeX genome diverges from Nipponbare to such an extent that no variants can be confidently identified. These appear either as gaps in coverage or as regions containing a concentration of natural variations between KitaakeX and Nipponbare. We can now use the KitaakeX sequence as the direct reference genome and detect mutations in highly variable regions. This approach will simplify analysis and increase confidence in the identification of FN-induced mutations. Because there are only 219 small variations between KitaakeX and Kitaake (Additional file 8), the KitaakeX genome can also be used as the reference genome for Kitaake.

Conclusions

The de novo assembly of the KitaakeX genome serves as a useful reference genome for the model rice variety Kitaake and will facilitate investigations into the genetic

basis of diverse traits critical for rice biology and genetic improvement.

Methods

Plant materials and growth conditions

Dr. Thomas W. Okita from Washington State University provided the Kitaake seeds, which were originally obtained from Dr. Hiroyuki Ito, Akita National College of Technology, Japan. Dr. Jan E. Leach at Colorado State University provided seeds for Zhenshan 97, Minghui 63, IR64 and 93–11. Seeds of Kasalath were provided by the USDA Dale Bumpers National Rice Research Center, Stuttgart, Arkansas. Seeds were germinated on 1/2x MS (Murashige and Skoog) medium. Seedlings were transferred to a greenhouse and planted 3 plants/pot during the springtime (Mar. 2, 2017) in Davis, California. The light intensity was set at approximately 250 μmol m^{-2} s^{-1}. The day/night period was set to 14/10 h, and the temperature was set between 28 and 30 °C [29]. Rice plants were grown in sandy soil supplemented with nutrient water. The day when the first panicle of the plant emerged was recorded as the heading date for that plant. Kasalath seeds were received later, and the heading date was recorded in the same way. The experiment was repeated in winter.

Construction of a phylogenetic tree

We obtained 178,496 evenly distributed SNPs by dividing the genome into 3.8 kb bins and selecting one or two SNPs per bin randomly according to the SNP density of the bin. Genotypes of all the rice accessions, including 3010 accessions of the 3 K Rice Genomes Project and additional noted accessions, were fetched from the SNP database RiceVarMap v2.0 [30] and related genomic data [31] and used to calculate an IBS distance matrix which was then applied to construct a phylogenetic tree by the unweighted neighbor-joining method, implemented in the R package APE [32]. Branches of the phylogenetic tree were colored according to the classification of the 3010 rice accessions [2].

Genome sequencing and assembly

High molecular weight DNA from young leaves of KitaakeX was isolated and used in sequencing. See (Additional file 1) for further details.

Annotation of protein-coding genes

To obtain high-quality annotations, we performed high throughput RNA-seq analysis of libraries from diverse rice tissues (leaf, stem, panicle, and root). Approximately 683 million pairs of 2 × 151 paired-end RNA-seq reads were obtained and assembled using a comprehensive pipeline PERTRAN (unpublished). Gene models were predicted by combining ab initio gene prediction, protein-based homology searches, experimentally cloned cDNAs/expressed-sequence tags (ESTs) and assembled transcripts from the RNA-seq data. Gene functions were further annotated according to the best-matched proteins from the SwissProt and TrEMBL databases [33] using BLASTP (E value $< 10^{-5}$) (Additional file 11). Genes without hits in these databases were annotated as "hypothetical proteins". Gene Ontology (GO) [34] term assignments and protein domains and motifs were extracted with InterPro [35]. Pathway analysis was derived from the best-match eukaryotic protein in the Kyoto encyclopedia of genes and genomes (KEGG) database [36] using BLASTP (E value$< 1.0e^{-10}$).

Genome Synteny

We used SynMap (CoGe, www.genomevolution.org) to identify collinearity blocks using homologous CDS pairs with parameters according to Daccord et al. [37] and visualized collinearity blocks using Circos [38].

Repeat annotation

The fraction of transposable elements and repeated sequences in the assembly was obtained merging the output of RepeatMasker (http://www.repeatmasker.org/, v. 3.3.0) and Blaster (a component of the REPET package) [39]. The two programs were run using nucleotide libraries (PReDa and RepeatExplorer) from RiTE-db [40] and an in-house curated collection of transposable element (TE) proteins, respectively. Reconciliation of masked repeats was carried out using custom Perl scripts and formatted in gff3 files. Infernal [41] was adopted to identify non-coding RNAs (ncRNAs) using the Rfam library Rfam.cm.12.2 [42]. Results with scores lower than the family-specific gathering threshold were removed; when loci on both strands were predicted, only the hit with the highest score was kept. Transfer RNAs were also predicted using tRNAscan-SE [43] at default parameters. Repeat density was calculated from the file that contains the reconciled annotation (Additional file 10).

Analysis of genomic variations

Analysis of SNPs and InDels: We used MUMmer (version 3.23) [26] to align the Nipponbare and Zhenshan97 genomes to the KitaakeX genome using parameters -maxmatch -c 90 -l 40. To filter the alignment results, we used the delta -filter – 1 parameter with the one-to-one alignment block option. To identify SNPs and InDels we used show-snp option with parameter (–Clr TH). We used snpEff [44] to annotate the effects of SNPs and InDels. Distribution of SNPs and InDels along the KitaakeX genome was visualized using Circos [38].

Analysis of PAVs and Inversions: We used the show-coords option of MUMmer (version 3.23) with parameters -TrHcl to identify gap regions and PAVs above 86

bp in size from the alignment blocks. We used the inverted alignment blocks with ≥98% identity from the show-coords output file to identify inversions.

To identify genomic variations between Kitaake and KitaakeX we sequenced and compared the sequences using the established pipeline [15].

BAC library construction

Arrayed BAC libraries were constructed using established protocols [45]. Please see Additional file 1 for further details.

Genome size estimation

We used the following methodology to estimate KitaakeX genome size:

(1) Using the Illumina fragment library, we created a histogram of 24mer frequencies. This was performed by first counting the frequency of all 24mers. The number of kmers at each frequency was tallied, and a histogram was created. (2) The kmer histogram generally indicates a peak value at a particular frequency corresponding to the average coverage of 24mers on the genome. (3) We then took the peak value representing the coverage on the genome, and computed the total bases in the Illumina library. Further dividing the total bases by the coverage, provided an estimate of the genome size. This value is generally accurate to +/− 10%.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6262-4.

Additional file 1: Supplementary Methods: Sequencing, genome assembly and construction of pseudomolecule chromosomes, and BAC library construction. **Table S1.** Genomic libraries included in the *Oryza sativa* (KitaakeX) genome assembly and their respective assembled sequence coverage levels in the final release. **Table S2**. PACBIO library statistics for single pass yield of the 42 chips included in the *Oryza sativa* (KitaakeX) genome assembly and their respective assembled sequence coverage levels. **Table S3.** Summary statistics of the output of the SuperNova whole genome shotgun assembly prior to integration with the PACBIO assembly. **Table S4**. Summary statistics of the raw output of the MECAT whole genome shotgun assembly. **Figure S1-S12.** Syntenic Japonica sequence placements on the *Oryza sativa* (var. KitaakeX) chromosomes. Each figure shows one chromosome. **Table S5.** Final summary assembly statistics for chromosome scale assembly **Figure S13.** Dot plot of BAC clone 119,492 on a region of Chr_02. **Figure S14.** Dot plot of BAC clone 120,743 on a region of Chr_12. **Figure S15.** Dot plot of BAC clone 119,503 in a region of Chr_06. **Table S6.** KitaakeX BAC libraries used for genome assembly and construction of pseudomolecule chromosomes. For Figures S1-S12, plot of the marker placements for each chromosome is shown.

Additional file 2: Table S7. BUSCO analysis of KitaakeX and comparison with other rice genomes. **Table S8.** Summary of transposable elements in KitaakeX, Nipponbare, and Zhenshan97. **Table S9.** Comparison of SNPs and INDELs between three rice genomes. **Table S10.** Comparison of single base substitutions between three rice genomes. **Table S11.** *Oryza sativa* KitaakeX annotation v3.1 on assembly v3.0. **Table S12**. Sequence length of pseudomolecules, number of genes and gene models for each of the 12 rice chromosomes.

Additional file 3: Table S13. Genes used in annotation quality control. We selected 291 genes from three pathways associated with stress resistance, flowering time and response to light to evaluate the quality of annotation. See main text for additional details.

Additional file 4. Comparative genomic analysis between KitaakeX and Nipponbare. SNPs, InDels, PAVs, Inversions, and genes affected by SNPs, IndDels, PAVs and Inversions are listed in this file.

Additional file 5. Comparative genomic analysis between KitaakeX and Zhenshan97. SNPs, InDels, PAVs, Inversions, and genes affected by SNPs, IndDels, PAVs and Inversions are listed in this file.

Additional file 6. SNPs between KitaakeX and Zhenshan97.

Additional file 7: Figure S16. Genomic variation showing gene variations between KitaakeX and Nipponbare and ZS97. ***a.*** Length distribution of InDels in protein-coding regions. ***b.*** SNPs and InDels that cause high-impact gene variations between KitaakeX and Nipponbare and ZS97. ***c.*** Gene enrichment in KitaakeX unique present regions compared with Nipponbare.

Additional file 8. Genomic variations between KitaakeX and Kitaake. SNPs, InDels variations, and XA21 position are listed in this file.

Additional file 9: Figure S17. Integrative genomics viewer (IGV) snapshot showing presence of XA21 transgene and selectable marker encoding a hygromycin B phosphotransferase on chromosome 6 of KitaakeX.

Additional file 10. Repeat annotation of KitaakeX genome.

Additional file 11. Functional annotation of KitaakeX genome.

Abbreviations

BAC: Bacterial Artificial Chromosome; BLAST: Basic Local Alignment Search Tool; BLAT: BLAST-like alignment tool; BUSCO: Benchmarking Universal Single-Copy Orthologs; EST: Expressed-Sequence Tags; FN: Fast Neutron; GO: Gene Ontology; KEGG: Kyoto encyclopedia of genes and genomes; MECAT: Mapping, Error Correction and de novo Assembly Tools; MS: Murashige and Skoog; NB-ARC: Nucleotide-Binding Adaptor shared by APAF-1, R proteins, and CED-4; PAVs: Presence/Absence Variations; SNP: Single Nucleotide Polymorphisms; TEs: Transposable Elements

Acknowledgements

We thank Rick A. Rios, Maria E. Hernandez, and Natasha Brown for assistance in genomic DNA isolation and submission and seed organization.

Authors' contributions

RJ, GL, MC and PCR conceived and initiated the study. RJ and GL carried out sequencing, assembly and annotation in collaboration with JJ, SS, DAK, JT, DD, DB, JG, DC, KWB, RAW, NTP, and JSRJ and GL conducted comparative genomics analyses. FX and WX contributed to the phylogenetic tree. RJ, PCR, GL and CRN wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by NIH (GM59962), NIH (GM122968) and NSF (IOS-1237975) grants to PCR. The work conducted by the Joint BioEnergy Institute and the Joint Genome Institute which was supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

Availability of data and materials

The genome sequencing reads and assembly have been deposited under GenBank under accession number PRJNA234782 and PRJNA448171 respectively. The assembly and annotation of the Kitaake genome are available at Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). The RNA-Seq reads of KitaakeX leaf, panicle, stem and root have been deposited under GenBank accession numbers SRP182736, SRP182738, SRP182741, and SRP182737 respectively. Genome sequencing reads for Kitaake have been deposited under GenBank under accession number SRP193308.

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no conflict of interest.

Author details
[1]Department of Plant Pathology and the Genome Center, University of California, One Shields Avenue, Davis, CA 95616, USA. [2]Feedstocks Division, Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]U.S. Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA. [4]HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA. [5]The Provincial Key Lab of Plant Pathology of Hubei Province and College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, Hubei, China. [6]National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China. [7]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA. [8]BIO5 Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA. [9]International Rice Research Institute, Genetic Resource Center, Los Baños, Laguna, Philippines. [10]Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland. [11]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.

Received: 7 June 2019 Accepted: 5 November 2019
Published online: 27 November 2019

References

1. Gross BL, Zhao Z. Archaeological and genetic insights into the origins of domesticated rice. Proc Natl Acad Sci U S A. 2014;111(17):6190–7.
2. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang FJN. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557(7703):43.
3. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y). 2013;6(1):4.
4. Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang XJ. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science. 2002;296(5565):79–92.
5. Schatz MC, Maron LG, Stein JC, Wences AH, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban EJG. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. Genome Biol. 2014;15(11):506.
6. Zhang J, Chen L-L, Xing F, Kudrna DA, Yao W, Copetti D, Mu T, Li W, Song J-M, Xie W *et al*: Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. 2016, **113**(35):E5163-E5171.
7. Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, Ma B, Qi M, Li Y, Zhao XJN. Sequencing and de novo assembly of a near complete indica rice genome. Nat Commun. 2017;8:15324.
8. Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, Zuccolo A, Song X, Kudrna D, Ammiraju JS, et al. The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. Nat Genet. 2014;46(9):982–8.
9. Kim SL, Choi M, Jung KH, An G. Analysis of the early-flowering mechanisms and generation of T-DNA tagging lines in Kitaake, a model rice cultivar. J Exp Bot. 2013;64(14):4169–82.
10. Ichitani K, Okumoto Y, Tanisaka T. Photoperiod sensitivity gene of se-1 locus found in photoperiod insensitive rice cultivars of the northern limit region of rice cultivation. Breed Sci. 1997;47:8.
11. Kunihiro Y, Ebe Y, Wada S, Shinbashi N, Honma A, Sasaki T, Sasaki K, Numao Y, Morimura K, Tan No H. The new rice variety Kita-ake. Bulletin of Hokkaido prefectural agricultural experiment stations. 1989;59:4.
12. Jung KH, An G, Ronald PC. Towards a better bowl of rice: assigning function to tens of thousands of rice genes. Nat Rev Genet. 2008;9(2):91–101.
13. Wang L, Zheng J, Luo Y, Xu T, Zhang Q, Zhang L, Xu M, Wan J, Wang MB, Zhang CJPJ. Construction of a genomewide RNA i mutant library in rice. Plant Biotechnol J. 2013;11(8):997–1005.
14. Gao H, Zheng XM, Fei G, Chen J, Jin M, Ren Y, Wu W, Zhou K, Sheng P, Zhou F, et al. Ehd4 encodes a novel and Oryza-genus-specific regulator of photoperiodic flowering in rice. PLoS Genet. 2013;9(2):e1003281.
15. Li G, Jain R, Chern M, Pham NT, Martin JA, Wei T, Schackwitz WS, Lipzen AM, Duong PQ, Jones KC, et al. The sequences of 1504 mutants in the model Rice variety Kitaake facilitate rapid functional genomic studies. Plant Cell. 2017;29(6):1218–31.
16. Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, Gardner J, Wang B, Zhai WX, Zhu LH, et al. A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. Science. 1995;270(5243):1804–6.
17. Gao H, Jin M, Zheng X-M, Chen J, Yuan D, Xin Y, Wang M, Huang D, Zhang Z, Zhou KJPNAS. Days to heading 7, a major quantitative locus determining photoperiod sensitivity and regional adaptation in rice. Proc Natl Acad Sci U S A, 16342. 2014;111(46):16337.
18. Ronald PC, Beutler B. Plant and animal sensors of conserved microbial signatures. Science. 2010;330(6007):1061–4.
19. Liu Y, Wu H, Chen H, Liu Y, He J, Kang H, Sun Z, Pan G, Wang Q, Hu JJN. A gene cluster encoding lectin receptor kinases confers broad-spectrum and durable insect resistance in rice. Nat Biotechnol. 2015;33(3):301.
20. Zhou X, Liao H, Chern M, Yin J, Chen Y, Wang J, Zhu X, Chen Z, Yuan C, Zhao W, et al. Loss of function of a rice TPR-domain RNA-binding protein confers broad-spectrum disease resistance. Proc Natl Acad Sci U S A. 2018;115(12):3174–9.
21. Rodrigues JA, Ruan R, Nishimura T, Sharma MK, Sharma R, Ronald PC, Fischer RL, Zilberman D. Imprinted expression of genes and small RNA is associated with localized hypomethylation of the maternal genome in rice endosperm. Proc Natl Acad Sci U S A. 2013;110(19):7934–9.
22. Li T, Liu B, Spalding MH, Weeks DP, Yang B. High-efficiency TALEN-based gene editing produces disease-resistant rice. Nat Biotechnol. 2012;30(5):390–2.
23. Xie K, Minkenberg B, Yang Y. Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. Proc Natl Acad Sci U S A. 2015;112(11):3570–5.
24. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
25. Guyot R, Keller BJG: Ancestral genome duplication in rice. 2004, 47(3):610–614.
26. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SLJG. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
27. Matsumoto T, Wu J, Itoh T, Numa H, Antonio B, Sasaki T. The Nipponbare genome and the next-generation of rice genomics research in Japan. Rice (New York, NY). 2016;9(1):33.
28. Tang X, Liu G, Zhou J, Ren Q, You Q, Tian L, Xin X, Zhong Z, Liu B, Zheng XJG. A large-scale whole-genome sequencing analysis reveals highly specific genome editing by both Cas9 and Cpf1 (Cas12a) nucleases in rice. Genome Biol. 2018;19(1):84.
29. Schwessinger B, Bahar O, Thomas N, Holton N, Nekrasov V, Ruan D, Canlas PE, Daudi A, Petzold CJ, Singan VR, et al. Transgenic expression of the dicotyledonous pattern recognition receptor EFR in rice leads to ligand-dependent activation of defense responses. PLoS Pathog. 2015;11(3):e1004809.
30. Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, Xing Y, Chen L, Xie W. RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Res. 2015;43(Database issue):D1018–22.
31. Li G, Chern M, Jain R, Martin JA, Schackwitz WS, Jiang L, Vega-Sanchez ME, Lipzen AM, Barry KW, Schmutz J, et al. Genome-wide sequencing of 41 Rice (Oryza sativa L.) mutated lines reveals diverse mutations induced by fast-neutron irradiation. Mol Plant. 2016;9(7):1078–81.
32. Paradis E, Claude J, Strimmer K. APE: analyses of Phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–90.
33. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28(1):45–8.
34. Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T. A gene ontology inferred from molecular networks. Nat Biotechnol. 2013;31(1):38–45.
35. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, et al. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 2017;45(D1):D190–9.
36. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61.
37. Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, van de Geest H, Bianco L, Micheletti D, Velasco R, et al. High-quality de novo

assembly of the apple genome and methylome dynamics of early fruit development. Nat Genet. 2017;49(7):1099–106.
38. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.
39. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011; 6(1):e16526.
40. Copetti D, Zhang J, El Baidouri M, Gao D, Wang J, Barghini E, Cossu RM, Angelova A, Maldonado LC, Roffler S, et al. RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. BMC Genomics. 2015;16:538.
41. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5.
42. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43(Database issue):D130–7.
43. Schattner P, Brooks AN, Lowe TM: The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 2005, 33(Web Server issue):W686–W689.
44. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92.
45. Luo M, Wing RA: An Improved Method for Plant BAC Library Construction. In: Plant Functional Genomics. Edited by Grotewold E. Totowa, NJ: Humana Press; 2003: 3–19.

Publisher's Note